# OLAP on Information Networks: a new Framework for Dealing with Bibliographic Data

Wararat Jakawat, Cécile Favre and Sabine Loudcher

Université de Lyon (ERIC LYON 2), France
{wararat.jakawat, cecile.favre, sabine.loudcher}@univ-lyon2.fr

**Abstract.** In the context of decision making, data warehouses support OLAP technology and they have been very useful for efficient analysis onto structured data. For several years, OLAP is also used to analyze and visualize more complex data. Now, many data sets of interest can be described as a linked collection of interrelated objects. They could be represented as heterogeneous information networks, in which there are multiple object and link types. In this paper, we are focusing on bibliographic data. This type of data constitutes a rich source that is the starting point of research on bibliometrics, scientometrics domains. In this context, we discuss the interest of combining information networks, OLAP and data mining technologies. We propose a framework to materialize this combination and discuss the main challenges to build this framework. The basic idea is to be able to analyze various networks built from the bibliographic data representing different points of view (authors networks, citations networks...) and their dynamic.

**Keywords:** OLAP, Data Warehouse, Information Networks, Bibliographic Data, Data Mining

## 1 Introduction

Communication systems, biological networks, transport systems, social and information systems on the web have become ubiquitous and their volume has increased every day. All these systems are networked systems and they usually consist of a large number of interacting and multi-typed objects [6]. Individual objects interact with a specific set of objects, forming large data sets, interconnected among them. Such interconnected, multi-typed networks or systems are called heterogeneous information networks [6, 12]. They are extracted from the web, blogs and various kinds of online databases. For example, social networks are extracted from postings and blogs like Facebook; highway networks are extracted from transportation databases; publication author networks and citation networks are extracted from bibliographic databases like DBLP and MedPub etc.

Graphs have been widely used for modeling these networks and there have been numerous studies dealing with information networks, in many disciplines. The goal is to understand the structure and the behavior of information networks. Extracting knowledge inside large networks is a time-consuming and

complex task. Problems including ranking, clustering, classification, entity similarity search and relationship prediction in information networks have been studied [14]. Extracting knowledge from an information network could answer questions such as *what are the main topics of a set of publications?, who are the central entities in a community ?*, etc. Moreover, with such knowledge, it is possible to understand past events and to predict events in future.

In parallel, data warehouses and OLAP (Online Analytical Processing) could be very useful for dealing with heterogeneous information networks. Data warehouse systems support OLAP or multidimensional data analysis by building cubes to provide easy navigation, visualization and fast analysis for decision making within a vast amount of data. Users can view data through several dimensions or analysis axis and through different hierarchical levels for each dimension via OLAP operators.

In this paper, we outline some actual researches about OLAP on information networks and we present a new framework. In our framework, we want to build several networks of a given study, these networks representing different points of view of a same problem dealing with bibliographic data. Our goal is to model and build multiple networks and then to store them into a data warehouse. After that, we want to use OLAP for visualizing and analyzing networks. Besides we plan to combine OLAP and some data mining techniques in order to enrich the network analysis. In this paper we address the issues of such a new framework considering the case of scientific bibliographic data. We chose to deal with bibliographic data as a first application domain to test our ideas. This is a position paper to discuss the basis for future work.

The remainder of this paper is organized as follows. Section 2 deals with bibliographic data and their interest for different approaches. Section 3 introduces concepts about information networks and OLAP. Section 4 outlines general definitions of OLAP on information networks and related work. Section 5 presents our proposed framework and the related challenges. Section 6 is a conclusion.

## 2 Bibliographic data

Bibliographic data analysis can be applied in many works in different areas. There are several objectives, including not only research evaluation, but also research evolution understanding, bibliometric analysis, etc. It could be useful for helping governments, managers and others to make their task easier such as deciding which projects or researchers should receive more support, who should be a reviewer, how to make evolve the topics of a conference or a journal over time.

Bibliographic data are extracted from online databases such as DBLP, ACM, PubMed, NCBI and etc. They collect large data about scientific publications in different domains, including information about authors (e.g. name and institutions) and details of publications (titles, conferences, keywords, published date

and citations). It is thus possible to build networks such as co-authors network, citations network and so on.

The network can be represented as a graph containing nodes and edges. For example, co-authors network contains authors as nodes and co-author relationship as edges.

Bibliographic data have been used as a basic of many studies focusing on different challenges. Muhlenbach *et al.* proposed to discover research communities [10]. They proposed a graph-based clustering method in the case of conferences and authors. Different kinds of relationships are considered. Gupta *et al.* designed a clustering algorithm for network evolution [5]. Their node types in the network were papers, authors, conferences and terms. The algorithm can take into account the evolution both at the object level and at the clustering level. Huang *et al.* introduced the detection of the evolution of semantic communities extracted from article titles [7]. They constructed a word association network based on word relationships in titles. They used statistical distribution frequencies on edges to classify two communities. Deng *et al.* presented three models of expert-finding approaches considering the publications [4]. Their models included the statistic language model, the topic-based model and a hybrid model. Pham and Klamma provided a visualization using citation analysis [11]. Social Network Analysis (SNA) is used to determine clustering issues. The result is presented on clustering level. Several researches studied the databases of published papers in order to provide a tool or a user interface for monitoring and exploring these data [9, 13].

## 3 Preliminaries

### 3.1 Information networks

An information network is a large number of individual objects interacting with a specific group of objects [3, 15]. Usually, an information network is visualized with a graph model. Each node represents an object or an entity such as actors in social networks, an edge or a link is a relationship between two entities.

**Definition 1.** *A graph G = (V,E) consists of V, a set of vertices or nodes and E, a set of edges. Each edge has two vertices associated with it.*

There are two types of networks. In the first type, networks are homogeneous networks. They contain a single object type and a single link type such as friends networks, authors networks and movies networks. Links may include a label or a weight. In the other type, networks are composed of multiple object and link types and they are called heterogeneous networks. For example, a medical network can contain patients, doctors, disease entities and links can be "*is followed by*" or "*has contracted*". The figure 1 shows two examples of bibliographic networks. In figure 1a the authors network is a homogeneous network where each node represents an author (*authorID*) and an edge represents a co-author relationship in one or several papers. For example, authors A and B have written

three papers together in the same conference. A link with the weight 3 has been added between them. An example of a heterogeneous network is presented in figure 1b, it is an author-paper network. This network has two types of nodes: authors and papers. There are three types of edges. The first link is "*written*" between authors and papers. The second represents co-author relationship and the last one relates papers written by the same author.


**Fig. 1.** Examples of bibliographic networks

### 3.2 On-line Analytical Processing (OLAP)

In data warehouse systems, On-line Analytical Processing (OLAP) gives a multidimensional view of data by building data cubes [2]. The multidimensional model consists of facts representing by measures and dimensions. The data cube contains cells that include measures, which are values based on a set of dimensions. Dimensions can be seen as analysis axis and may be organized into hierarchies with several levels. Levels are structured attributes or not. For instance, in the example of the publications, the *time* dimension hierarchy may consist of four levels: *semester*, *year*, *decade*, *all*; the *venue* dimension hierarchy includes three levels: *support* (the name of the conference like *ICDM*, the name of the journal like *TKDE*, the name of the book, etc.), *research area* (like *databases*, *data mining*, *information retrieval*, etc.) and *all*. The dimensions are assumed to determine measures. Basically, measures can be numerical indicators which are calculated by aggregating the same dimensions of all facts.

An interestingly feature of the multidimensional model is the measure aggregation by using one or more dimensions, e.g., computing the total number of publications by each country over years. There are four classic OLAP operations: *roll-up* takes the current data and does a group-by on one dimension in order to aggregate or summarize facts; *drill-down* is the dual of the roll-up operator by giving more details; *slice and dice* reduce dimensions for taking a subset of data on its dimensions and *pivot* changes layouts for analyzing in different points of view.

## 4 OLAP on information networks

### 4.1 General definitions

First, Chen *et al.* introduced Graph OLAP, a general framework for OLAP on information networks [3]. Graph OLAP is a collection of network snapshots where each snapshot $i$ has $k$ informational attributes describing the snapshot and has a graph $Gi = (Vi, Ei)$. Such snapshots represent different sets of the same objects in real applications. For instance, with regard to the author-paper network of the figure 1b, *venue* and *time* informational attributes can mark the status of each individual snapshot e.g. *ICDM 2008* and *ASONAM 2010*; *authorID* is a node attribute defining each node, and collaboration frequency is an edge attribute reflecting the connection strength of each edge. Dimension and measure concepts, found in traditional OLAP domain, should be re-defined for Graph OLAP.

At first, there are actually two types of graph OLAP dimensions. The first one is an informational dimension, and it uses an informational attribute. These dimensions have two roles: organizing snapshots into groups based on different perspectives and granularity (each group corresponds to a cell in the OLAP cube) and controlling snapshot views but they do not touch the inside of any individual snapshot. For example, the two informational attributes *venue* and *time* with their respectively hierarchical concepts {*semester*, *year*, *decade*, *all*} and {*support*, *research area*, *all*} can be used as informational dimensions. We can look at the snapshot of each group e.g., (*ICDM*, *all years*) and (*data mining area*, *2010*).

The second type of dimension is a topological dimension coming from the attributes of topological elements. Topological dimensions operate on nodes and edges within individual networks. Let us consider author network for instance, the following hierarchy {*institute*, *country*, *continent*, *all*} associated with the node attribute *authorID* can be used for merging authors from a same institute into a generalized node. A new graph with generalized nodes is generated by summarizing the original network. In our example it shows interactions among institutions.

There are two kinds of measures in Graph OLAP. The first one is a graph. Graph is both viewed as a data source and as a special kind of measures. The second kind of measure is not a graph. It could be a node count, average degree, centrality etc. Due to different types of dimensions in graph OLAP, there are different semantics for aggregation. Let us consider an aggregated graph measure for example, aggregating data with informational dimensions groups among the snapshots such as collaborations between authors in the same conferences and during a period of time. Users can *roll-up* on the papers and grouped them by research areas. Whereas aggregating data with topological dimensions groups elements inside individual networks such as a new generalized network from au-

thor network is generated in order to have an institution network.

After this general framework proposed by Chen *et al.*, we propose a comparison between traditional OLAP and Graph OLAP (see table 1). Traditional data warehouses focus on the storage and data retrieval in contrast of data warehouses over graphs that are interested in representing information networks which are interrelated and multi-typed. Traditional data cubes take facts and generate aggregate measures. Graph cubes consider both attributes and structures for network aggregation. A given network as input is changed into a new network as output. Two types of dimension have been presented in Graph OLAP (informational and topological dimension) whereas there is only one type in traditional OLAP. In term of measures, traditional OLAP has numeric measures and aggregation functions such as COUNT and SUM to summarize multiple records. There are two types of measures in Graph OLAP. First, the measure can take the form of a graph and the aggregation function is then specific to graph. The second type of measure is not graph but can be indicators coming from graph theory such as average degree and diameter.

In traditional OLAP there is only one semantic for operators such as roll-up. The OLAP semantics accomplished through informational dimensions and topological dimensions are different and Chen *et al.* speak about informational OLAP (abbr. I-OLAP) and topological OLAP (abbr. T-OLAP), respectively. With roll-up in informational OLAP, snapshots are just different observations of the same underlying network, and they are grouped into one cell in the cube, without changing the network structure. For roll-up in topological OLAP, networks are not grouped but the reorganization is inside individual snapshots and a new generalized graph is built with a new topological structure. Lastly, a traditional data warehouse does not consider relationships between records.

**Table 1.** Comparison between traditional OLAP and Graph OLAP

| | Traditional OLAP | Graph OLAP |
|---|---|---|
| Input | Facts in cuboids | A given network with snapshots |
| Output | Aggregated measures | A new network more generalized |
| Dimensions | Attributes | Informational and topological |
| Hierarchies | Yes | Yes (both for info. and topo. dimensions) |
| Measures | Numeric indicators | Aggregated graph measure |
| | | Measures coming from graph theory |
| | Aggregation function (count, sum, average) | Specific aggregation functions |
| Operations | Roll-up, drill-down, slice & dice, pivot | Operations within informational or topological OLAP |
| Problems | Not considering links among data records | How taking interactions among entities into account |

### 4.2 Literature review

In recent years, many researchers have been interested in OLAP on information networks. Wei proposed a concept of link OLAP based on link-oriented analysis [17]. It extended entity analysis to link analysis. However, he did not propose new models or operations. Tian *et al.* introduced an operation called *SNAP* [16]. It can produce a summary graph by grouping nodes. Moreover, users can control the different resolutions of summaries by a k-SNAP operation. Chen *et al.* and Qu *et al.* proposed a data cube on graphs [3, 15]. Chen's framework used the top-10 central method for visualization over the data cube. While Qu's proposal efficiently computed measures and user's requests with two measure properties: T-Distributiveness and T- Monotonicity. Zhao *et al.* introduced a new data warehouse model [19]. This model is called *Graph cube* and it supports a new class of user queries called *crossboid*. Their model considered network aggregation both on entities and relationships. Kampgem *et al.* presented a mapping from linked data to data cube [8]. They integrated statistic linked data into format for loading into OLAP systems. Trifonova *et al.* presented an application for analyzing bibliographic networks [13] and they used star schema to design the data warehouse. It allowed data extraction from data cubes and authors using tabular and graphical views. Yin *et al.* defined a concept of entity dimensions to support two dimensions of heterogeneous networks [18]. They proposed *HMGraph OLAP* a data warehouse model using constellation schema. They designed novel operations named *Rotate* and *Stretch*. The previous studies did not mention how to improve performances on cube materialization. Therefore, Yin's approach demonstrated the strategy of index graphs. Researches are interested in studies about OLAP on information networks based on multidimensional and multilevel concepts. All of them have not provided a query language to support n-dimensional computations on graph OLAP. So Beheshti *et al.* proposed a graph data model and a query language extended SPARQL [1]. Their model considered both objects and links among them and there are two kinds of dimensions of information networks.

## 5 Proposed framework

The previous works have been interested in the effectiveness and efficiency of Graph OLAP to provide OLAP on information networks. The major limitation of these studies is that building a data warehouse is limited to only one network.

Our proposal is to design a data warehouse and OLAP analysis for several networks. The proposed framework is shown in figure 2. The starting point is databases of bibliographic data. We examine three online databases in computer science domain (DBLP, ACM and PASCAL) that allow us to collect publications under the form of XML data. Our idea is to build different networks from such databases: co-authors network, citations network, topics network, conferences network and so on. The networks are represented under the form of graphs. We would represent different actors and types of links as follows:

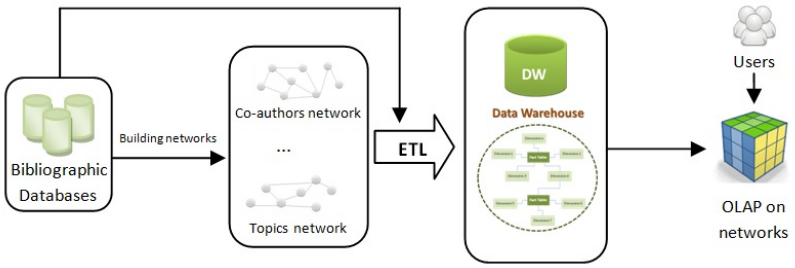


**Fig. 2.** The proposed framework

- co-author network is created with authors as vertices and co-author relationships as edges,
- in citation network, vertices are papers and edges represent a relationship between the cited and citing documents,
- topic network contains topic areas as vertices and the same area as relation,
- conference network contains names of conference as vertices and the same area as relation.

Many techniques can be used to extract knowledge from those networks such as data mining and SNA methods. The extracted knowledge can enrich networks. For example, clustering is useful to discover communities in many systems. It is to classify groups of entities that share similar properties and to provide the changes in the objects over time such as discovering organizational relations and identifying researcher communities. For instance, if we consider communities detection, the result could be used to enrich hierarchies with new levels of data. Ranking is to evaluate objects of networks based on mathematical or statistical functions. It needs to calculate the distance between objects and the cluster center. However, combining both clustering and ranking may lead to more better results. For instance, ranking authors related to conference cluster by using the number of citations, the most popular topics in each institutions and top-10 of researchers in research areas. SNA methods are used to study the relationships, analyze citations, compute communications and calculate indicators. For example, a concept of closeness is calculated as a relevant score for finding collaborators on similar topics. Degree centrality provides an answer to the question "who are the leaders among researchers or popular research topics?".

Next, networks are loaded into a data warehouse through ETL process (Extract, Transform and Load). Different models are used to represent different networks. The structure of a data warehouse should be designed, it is based on the multidimensional model. It should be able to store the different networks and the related extracted knowledge. The fact can be a single node (an author, an institution) or a network (co-publications network). In our knowledge, there are many types of measure. Firstly, the measure can be classical like a numerical feature such as the number of papers, the number of citations

and the number of downloads. The measure can be textual such as keywords. In social network analysis, the measure can be the centrality, the diameter or the similarity. Lastly, a network can be a measure in *Graph OLAP*. In term of the aggregation function, it depends on the types of measure and on hierarchy concept. With classical measures, the *SUM* or *AVERAGE* functions are well suitable, they can construct a group of authors by laboratory or institution. An other example in *Graph OLAP*, graph summarization is to cluster authors by relationships. In our framework, we want to have the several types of measure and the adapted aggregation functions.

At the end, we plan to create OLAP tools for network aggregation, visualization and navigation. We have to answer users' queries such as navigating within other authors in collaboration who work on the same research topics. For efficient visualization and for network aggregation, we want to take into account both attributes of nodes and links between nodes. More, we have to analyze the dynamic of a network (authors, publications) over time such as the most popular topics in each year. In order to create these new OLAP tools, we plan to combine data mining methods and OLAP operators.

Considering this proposed framework, we have identified several challenges. First, we have to build the several networks by extracting them from databases and we have to extract knowledge form networks in order to enrich them. For this double task, we have to consider the existing algorithms and data mining techniques would be very useful. Secondly, a big challenge is how to design the model for storing multi-networks and knowledge. We think that classical models cannot meet our needs and we probably led to invent a new model. Thirdly, we have to consider the ETL step. How to consider this phase both for networks and knowledge ? Last, there is a crucial challenge to provide analysis tools, dealing with the various considered networks. Innovative tools should be developed for users.

## 6 Conclusion

In this paper, we discussed the interest of combining OLAP technology and information networks in the context of bibliographic data analysis.

We presented a related work on the use of these two domains to emphasize how it is possible to combine them. We also proposed a tentative framework to analyze bibliographic data taking benefit from these two areas and we addressed the main challenges to solve. The main ideas are (i) building various networks from the bibliographic databases such as DBLP, ACM... (co-author network, citations network, topics network, conferences network) ; (ii) building a data warehouse with the appropriate model to explore these information ; (iii) applying data mining techniques to enrich this information (such as detecting communities to enrich dimension hierarchies of the data warehouse) ; and (iv) developing appropriate tools (inspired from OLAP navigation process) for visualizing these data. Various problems have to be solved, such as summarizability

and topological issues. From a technical point of view, we need to explore existing tools and their usage (for instance graph database tool such as neo4j[1]).

In terms of perspectives of this preliminary work, we aim at dealing with every underlined challenges to provide a complete solution implementing our framework that combines OLAP technology, data mining and information networks to deal with bibliographic data.

## References


1. Beheshti, S.M.R., Benatallah, B., Motahari-Nezhad, H.R.: A Framework and a Language for On-Line Analytical Processing on Graphs. X.S. Wang et al. (Eds.): WISE'12. LNCS 7651, pp. 213–227 (2012)
2. Chaudhuri, S., Dayal, U.: An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD. vol. 26, no. 1, pp. 65–74(1997)
3. Chen, C., Yan, X., Zhu, F., Han, J., Yu, P.S.: Graph OLAP: Towards online analytical processing on graphs. ICDM'08. pp. 103–112(2008)
4. Deng, H., King, I., Lyu, M.R.: Formal Models for Expert Finding on DBLP Bibliography Data. ICDM'08. pp. 163–172 (2008)
5. Gupta, M., Aggarwal, C.C., Han, J., Sun, Y.: Evolutionary Clustering and Analysis of Bibliographic Networks. ASONAM'11. pp.63–70 (2011)
6. Han, J.: Mining Heterogeneous Information Networks by Exploring the Power of Links. DS'09. pp. 13–30 (2009)
7. Huang, Z., Yan, Y., Qiu, Y., Qiao, S.: Exploring Emergent Semantic Communities from DBLP Bibliography Database. ASONAM'09. pp. 219–214 (2009)
8. Kampgen, B., Harth, A: Transforming statistical linked data for use in OLAP systems. I-SEMANTICS. pp. 33–40 (2011)
9. Klink, S., Reuther, P., Weber, A., Walter, B., Ley, M.: Analysing Social Networks Within Bibliographical Data. DEXA'06. LNCS 4080, pp. 234–243 (2006)
10. Muhlenbach, F., Lallich, S.: Discovering Research Communities by Clustering Bibliographical Data. WI-IAT'10. vol. 1. pp. 500–507 (2009)
11. Pham, M. C., Klamma, R.: The Structure of the Computer Science Knowledge Network. ASONAM'10. pp. 17–24 (2010)
12. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: RankClus: integrating clustering with ranking for heterogeneous information network analysis. EDBT'09. pp. 565–576 (2009)
13. Trifonova, T. G.,: Warehousing and OLAP Analysis of Bibliographic Data. Intelligent Information Management. vol. 3, pp. 109–197 (2011)
14. Yu, P. S.: Information networks mining and analysis. APWeb'11. pp. 1–2 (2011)
15. Qu, Q., Zhu, F., Yan, X., Han, J., Yu, P.S., Li, H.: Efficient Topological OLAP on Information Networks. DASFAA'11. Part I. LNCS, vol. 6587, pp. 389–403 (2011)
16. Tian, Y., Hankins, R.A., Patel, L.M.: Efficient Aggregation for Graph Summarization. SIGMOD Conference. pp. 567–580 (2008)
17. Wei, W.: Complex network virtualization and link OLAP. (2007)
18. Yin, M., Wu, B., Aeng, Z,: HMGraph OLAP: a Novel Framework for Multidimensional Heterogeneous Network Analysis. DOLAP'12. pp. 137–144 (2012)
19. Zhao, P., Li, X., Xin, D., Han, J.: Graph cube: on warehousing and OLAP multidimensional networks. SIGMOD'11. pp. 853–864 (2011)


[1] www.neo4j.org